Journal of Heredity 2013:104(5):595–600
doi:10.1093/jhered/est050

The i5K Initiative: Advancing Arthropod Genomics for Knowledge, Human Health, Agriculture, and the Environment

i5K Consortium*

*Authors listed in the Appendix.

Address correspondence to Dr. Jay D. Evans, US Department of Agriculture, Agricultural Research Service, Bee Research Laboratory, Beltsville, MD 20705, or e-mail: jay.evans@ars.usda.gov.

Insects and their arthropod relatives including mites, spiders, and crustaceans play major roles in the world's terrestrial, aquatic, and marine ecosystems. Arthropods compete with humans for food and transmit devastating diseases. They also comprise the most diverse and successful branch of metazoan evolution, with millions of extant species. Here, we describe an international effort to guide arthropod genomic efforts, from species prioritization to methodology and informatics. The 5000 arthropod genomes initiative (i5K) community met formally in 2012 to discuss a roadmap for sequencing and analyzing 5000 high-priority arthropods and is continuing this effort via pilot projects, the development of standard operating procedures, and training of students and career scientists. With university, governmental, and industry support, the i5K Consortium aspires to deliver sequences and analytical tools for each of the arthropod branches and each of the species having beneficial and negative effects on humankind.

Key words: *comparative genomics, disease vector, agriculture, insect evolution, genome sequencing*

Humans share the planet with millions of arthropod species, only a handful of which depend on us as much as we depend on them. Arthropods include the well-known insects as well as spiders, mites, millipedes, and crustaceans such as crabs and shrimp. Insects alone include over 1 million described species and many-fold more remaining to be described. Arthropods are intimately tied to the earth's plants as pollinators and herbivores, and also play major roles in soil movement and nutrient cycles on land and in water. They compete with humans for crop plants and natural forage and benefit humans as mutualistic pollinators and alternate food sources. They transmit some of the most devastating diseases of humans, livestock, and plants yet have provided society with medical models for studying cancer, obesity, alcoholism, and neurological diseases. Beginning with the sequencing of the *Drosophila melanogaster* genome in 2000 (Adams et al. 2000) and aided by relatively small genome sizes for many of the best-studied species (Gregory 2013; Grbic et al. 2011), insects and now mites have provided general insights into the evolution of genomes and the use of genomic technologies to infer basic processes of evolution, development, physiology, reproduction, and survival.

The 5000 arthropod genomes initiative (i5K) reflects a diverse group of scientists addressing fundamental and applied biological questions. Participation in the i5K initiative is open and voluntary and demonstrates the success of coordinated efforts by scientists who are passionate about arthropods and the questions that can be answered through genomic and genetic analyses. With an entry-point wiki (http://arthropodgenomes.org/wiki/i5K) receiving over 60 000 hits by July, 2013 and growing communities centered on specific taxa, biological systems, biodiversity, and issues of human welfare, the i5K initiative has grown in the past 2 years to include 600 participants and more than 800 arthropod species nominated for genome sequencing. In this article, we describe our mission, highlight the first i5K Community Workshop, and review the various challenges and successes of this growing effort. We also advertise the current online resources for i5K in hopes that readers will join future efforts.

What Is the Mission of the i5K Initiative?

The i5K initiative is not a funding agency, nor does it have a single core group or institution that will tackle even a modest subset of the 5000 targeted species for genomic analyses. Instead, this initiative seeks to have an impact through 1) prioritizing and promoting genome projects, 2) providing guidelines for designing and carrying out an arthropod genome project, 3) training scientists in the tools needed and standards expected for genome projects, 4) fostering discussions that

reduce redundant efforts and enhance the impacts of genome projects, and 5) presenting recommendations to fundholders for the support needed to provide new opportunities for arthropod genomics in addressing fundamental research questions.

1st i5K Community Workshop

With support from the American Genetic Association and others and under the auspices of the Arthropod Genomics Consortium, the 1st i5K Community Workshop was held in Kansas City, MO, 29–30 May 2012. A diverse community of 140 participants from academia, government, and industry attended this workshop (Figure 1).

After an illuminating Plenary talk by O. Ryder (San Diego Zoo) describing parallel efforts in vertebrate species (Genome10K Community of Scientists 2009), the workshop featured research and forward-thinking presentations by scientists deeply interested in arthropods. Presentations by J. Coddington (Smithsonian Institution) and M. Pfrender (University of Notre Dame) highlighted the logistics of identifying key species on the tree of life, curating these species, and preparing material for sequencing and analyses. This was the first public report and opportunity for reflection on 1 year of intense discussions by the i5K Working Group charged with vetting species as a whole and choosing a core set that have extremely high priority for sequencing. Representatives from the US National Science Foundation and US Department of Agriculture's National Institute of Food and Agriculture (G. Gilchrist and M. Purcell-Miramontes, respectively) reviewed competitive programs that support genomic studies and possible avenues for coordinated research and training efforts. Finally, biologists and informaticists gave their insights into what makes a genome project successful and highlighted currently available tools for sequencing, assembling, and analyzing genomes.

The second day was devoted to breakout groups discussing "Funding and Resources," "Taxon and Sampling Decisions," "Sequencing and Assembly," and "Community Annotation, Curation, and Delivery." Discussions of logistics and routines for sequencing and assembling various arthropod genomes were extensive and it was evident that many researchers were moving ahead with generating sequences for their favored species. Standardized protocols and data pipelines were identified as pressing needs. These pipelines would be updated frequently and would present methods for minimizing genetic variation within source individuals, sample collection, sequencing, assembly, annotation, and comparative genomics. Several workshop participants led by S. Richards (Baylor College of Medicine) began work to develop this set of standard operating procedures for successful arthropod genome projects, and their work continues. Additional breakout groups from the workshop developed wiki nodes that will become open forums for discussion (http://arthropodgenomes.org/wiki/i5K_working_groups). As the workshop ended, i5K leader Dr. G. Robinson provided a bridge by delivering the Plenary introduction to the 6th Arthropod Genomics Symposium, a venue for discussing progress and tools in arthropod research.

Since the workshop, Dr. K. Hackett of the USDA's Agricultural Research Service has coordinated regular teleconference and data calls to discuss the questions posted above and to look for new avenues for funding and facilitating genome projects. The topics of these calls cover each of the i5K working groups, and an effort is made to bring in outside participants from genome centers and funding bodies. One exciting direct outcome of these discussions is a commitment by the Baylor College of Medicine's Human Genome Sequencing Center to sequence, assemble, and annotate 30 key arthropod species. These species include economic pests such bed bugs, crop pests including the medfly and Brown marmorated stink bug, and an aquatic amphipod widely used to assess environmental threats. Spiders, crustaceans, and the earliest branches of the insect lineage are also represented. Similarly, the Beijing Genome Institute has developed a set of candidate species based on community discussions, and scientists there are initiating roughly the same number of genome projects. The Baylor project is under way, and sequencing of nearly half the targeted species has already started. More details can be found at https://www.hgsc.bcm.edu/content/i5K-project).

Members of the i5K Consortium have given presentations and led discussions at each of the major international entomology meetings since inception and have joined forces with like-minded colleagues at meetings focused on animal genomics, insect evolution, and genomic tools (e.g., iPlant consortium; http://www.iplantcollaborative.org) to discuss joint efforts and synergies. Additionally, members of the i5K Consortium are involved in the development of tools for visualization and editing of genome annotations (e.g., Web Apollo; http://gmod.org/wiki/WebApollo), insect genome databases, and the development of shared computational resources.

Challenges

Choosing Arthropod Genome Projects with the Most Impact

Significant efforts have been put forward in the past year to evaluate and prioritize arthropod species for sequencing. These evaluations, driven in part by submissions to the i5K wiki (Table 1) and by conversations started during the 2012 i5K Workshop, have been finalized by a committee of arthropod systematists and geneticists (chaired by M. Pfrender from the University of Notre Dame), leading to rankings of candidate taxa that will be useful for fundholders and project consortia.

Several criteria were considered in evaluating candidate taxa, including ecological roles, human impacts, conservation needs, and intriguing biology. In addition, efforts have been made to identify suitable taxa representing each of the major arthropod groups. Candidates will also be ranked on the basis of feasibility, with large genome size, small physical size, and rarity as aggravating factors for some taxa. Experience at the Baylor College of Medicine (http://www.hgsc.bcm.tmc.edu/

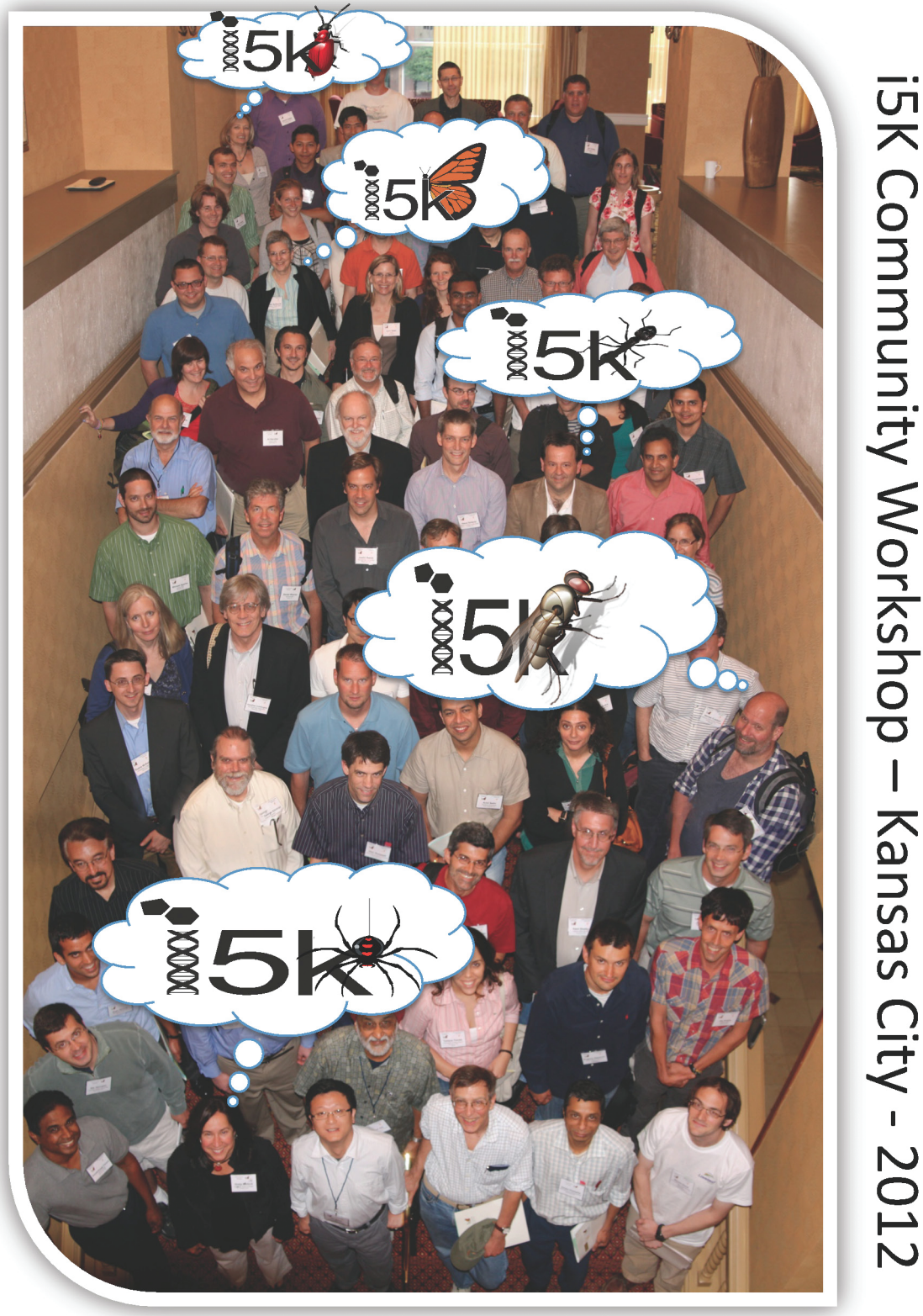

Figure 1. Participants in the i5K Community Workshop, expressing their interests in diverse taxa for sequencing.

Table 1 Nominated arthropod taxa for genome sequencing from the classes Hexapoda, Chelicerata, Crustacea, and Myriapoda, as part of the i5K initiative

Taxon group	Common names		Taxon group	Common names	
Hexapoda		***n* = 703**	**Chelicerata**		***n* = 63**
Hymenoptera	Ants, wasps, bees	256	Araneae	Spiders	34
Diptera	Flies	107	Opiliones	Daddy longlegs	3
Coleoptera	Beetles	69	Ixodida	Ticks	8
Lepidoptera	Moths, butterflies	56	Prostigmata	Mites	9
Strepsiptera		1	Astigmata	Mites	1
Siphonaptera	Fleas	3	Mesostigmata	Mites	2
Hemiptera	True bugs, aphids, planthoppers	91	Pantopoda	Sea spiders	1
Isoptera	Termites	25	Scorpiones	Scorpions	3
Orthoptera	Grasshoppers, crickets	15	Pseudoscorpiones	Pseudoscorpions	2
Blattaria	Roaches	7			
Dermaptera	Earwigs	4	**Crustacea**		***n* = 20**
Embioptera	Webspinners	2	Isopoda	Pill bugs	5
Ephemeroptera	Mayflies	2	Decapoda	Crabs, lobsters	5
Mantodea	Praying mantids	5	Amphipoda	Scuds, whale lice	2
Mecoptera	Scorpionflies	3	Euphausiacea	Krill	0
Megaloptera	Dobsonflies, alderflies	2	Diplostraca	Water fleas	0
Neuroptera	Lacewings, antliions	6	Anostraca	Fairy shrimps	1
Raphidioptera	Snakeflies	1	Notostraca	Tadpole shrimps	1
Notoptera	Grylloblattids, mantophasmatids	2	Brachypoda	Horseshoe shrimps	1
Odonata	Dragonflies, damselflies	3	Calanoida	Copepods	2
Phasmatodea	Stick insects	2	Harpacticoida	Copepods	1
Phthiraptera	Lice	2	Leptostraca	Sea fleas	1
Plecoptera	Stoneflies	2	Nectiopoda		1
Thysanoptera	Thrips	5			
Trichoptera	Caddisflies	4	**Myriapoda**		***n* = 6**
Protura		3	Chilopoda	Centipedes	2
Diplura		4	Diplopoda	Millipedes	2
Collembola	Springtails	10	Pauropoda		1
Archeognatha	Bristletails	5	Symphyla		1
Zoraptera		1			
Zygentoma	Silverfish	5			

content/i5K-project) and elsewhere has shown that genetic polymorphism within (often pooled) samples, quality of DNA or RNA, and poor vouchering or storage of samples significantly impair resulting genomic analyses. Attempts to mitigate these factors via novel sequencing and genome assembly methods are now being explored, with a hope that these techniques will allow projects for otherwise compelling species.

Providing a Roadmap for Genome Projects and Arthropod Resources

Researchers hoping to exploit genomic data to understand phenotypes, test biological hypotheses, or apply insights for pest control or the preservation of beneficial arthropods require guidance to find relevant sequence and metadata resources. Similarly, groups that are poised to start genome projects need current data on planned or in-hand genomic, transcriptomic, or metagenomic resources for their targets. Nominations at the i5K taxon wiki currently exceed 800 species, and these entries are serving a need by connecting community members and potentially competing groups focused on specific taxa.

Additionally, the inclusion of available genome size estimates makes it apparent that many taxa, among them the crickets and grasshoppers, will be challenging targets due to large or repetitive genome sequences. Moving forward, this community-driven list will continue to be populated with contributions from the i5K membership and with regular searches through large, institutional databases such as at the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/genome/?term=arthropoda) and the European Bioinformatics Institute (http://www.ebi.ac.uk/services). The expectation is that the i5K site and other resources will also serve to alert community members to genome projects on their favored organisms and provide a forum for early genome planning stages.

Populating Databases with Genome-Relevant Information for Each Species

Taxa with already completed genome sequences have received more extensive postsequencing analysis, curation, and database development than is likely to occur for upcoming genome candidates. At one extreme, sequence and phenotype data for numerous *Drosophila* species are

housed in FlyBase, a longstanding and constantly updated database (www.flybase.org). In addition, genomic resources for several crop pests are presented at Agripestbase (www.agripestbase.org), while arthropods that transmit disease to humans and other vertebrates are featured at VectorBase (www.vectorbase.org). Honey bees and other bees, wasps, and ants have found a stable home at the Hymenoptera Genome Database (www.hymenopteragenome.org), while flour beetles (BeetleBase; www.beetlebase.org), moths (i.e., the silkworm genome databases SilkDB, www.silkworm.genomics.org.cn/, and KAIKObase, sgp.dna.affrc.go.jp/KAIKObase/) and pea aphids (AphidBase, www.aphidbase.org) have current sites devoted to incoming genomic and/or biological data. In contrast, limited financial and informatics resources will preclude extensive project sites for many of the targeted i5K species. Nevertheless, the i5K Consortium agrees that single-page entries should be maintained for each of these species, to ensure permanence and availability of their resources. These pages are expected to provide a roadmap to finding pertinent sequence resources at NCBI and elsewhere, available literature, and phenotype data relevant to the species of interest. As one example, the USDA National Agricultural Library has developed postsequencing resources for genome projects in need of such resources (http://i5K.nal.usda.gov/).

Future Training and Education Needs

Since most scientists working with arthropods have little training in genomics and the exploitation of genomic resources, advances in this field will largely depend on the rate of workforce development. Workshops such as the annual Arthropod Genomes Conference and future i5K Workshops will initially fulfill this role, but long-term apprenticeships and training opportunities for entomology students and postgraduates in academic institutions are necessary to better equip individuals to meet the challenges of generating reliable genomic information. The Arthropod Genomes wiki will be one tool for describing and ranking such courses.

Since even the best-trained individuals are unlikely to implement or benefit from solo genome projects, there is a great need for building research communities of individuals around single and multispecies efforts. Taking advantage of division of labor by skill sets and interests, these communities will take shape around nearly all genome projects. Another goal of i5K is to act as a discussion board aimed at building such communities. To this end, in a U.S. National Science Foundation-sponsored Nescent workshop (www.nescent.org) this spring, i5K Community Curation group leaders A. Papanicolaou (Commonwealth Scientific and Industrial Research Organisation) and M. Muñoz-Torres (Lawrence Berkeley National Laboratory), brought together 12 software engineers and biologists to design and develop novel software for evolutionary genomics and genome annotation. i5K can also guide communities by ensuring that projects are built around a meritocracy (crediting the students and junior scientists who do much of the work) and are taking appropriate advantage of crowdsourcing and public resources.

How Can You Get Involved with i5K?

The first step to become involved with the i5K community is to request a user account at the wiki (http://www.arthropodgenomes.org/wiki/Special:RequestAccount), and then to develop a personal page with research interests. Users can subsequently expand the wiki and develop new pages for discussion, as well as contribute toward ongoing discussions and solicitations and join working groups. A key offshoot of this site will be the forging of new connections between individuals who are passionate about studying a specific taxon, as well as those eager to compare or annotate specific gene families or processes. An i5K initiative bioproject page at NCBI (http://www.ncbi.nlm.nih.gov/bioproject/165207) provides a means to link arthropod genome projects that are submitted to NCBI. In addition to the i5K pilot project at Baylor, organizers of independently supported projects involving bees, beetles, and other insects have linked these to the bioproject page, something we would encourage of all arthropod projects as part of NCBI data submission.

Conclusions

Although the i5K initiative is in its early stages, immediate needs have been identified and thus future directions have been drawn. Discussions spurred by the i5K Consortium have led community members to keep abreast of new technologies in sequencing, assembly, and annotation. Training possibilities are improving, both during i5K-sponsored workshops and via advertisements made by community members. Criteria for setting minimum requirements for pre-sequencing data and vouchering, as well as for needed landmarks for useful sequencing projects, are being actively developed. Methods to streamline the submission of genome projects into the NCBI Bioprojects schema, and to initiate public annotation tools, are being formalized. Finally, there is ongoing work toward supporting and expanding efforts to develop core ortholog sets for arthropods and to develop a controlled vocabulary for arthropod-relevant genetic traits.

Funding

American Genetic Association; the Arthropod Genomics Center (Kansas State University); and the US Department of Agriculture (National Institute for Food and Agriculture and Agricultural Research Service).

Appendix

Coauthors: Jay D. Evans (USDA-ARS, Beltsville, MD), Susan J. Brown (Kansas State University, Manhattan, KS), Kevin J. Hackett (USDA-ARS, Beltsville, MD; kevin.hackett@ars.usda.gov), Gene Robinson (University of Illinois, Urbana-Champaign, IL; generobi@illinois.edu), Stephen Richards

(Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX; stephenr@hgsc.bcm.edu), Daniel Lawson (European Bioinformatics Institute-Hinxton, United Kingdom; lawson@ebi.ac.uk), Christine Elsik (University of Missouri, Columbia, MO; elsikc@missouri.edu), Jonathan Coddington (Smithsonian Institution-NMNH, Washington, DC; coddington@si.edu), Owain Edwards (CSIRO, Centre for Environment and Life Sciences, Floreat, Australia; owain.edwards@csiro.au), Scott Emrich (University of Notre Dame, South Bend, IN; semrich@nd.edu), Toni Gabaldon (Centre for Genomic Regulation, Barcelona, Spain; toni.gabaldon@crg.es), Marian Goldsmith (University of Rhode Island, Providence, RI; mki101@uri.edu), Glenn Hanes (USDA-ARS, Beltsville, MD; Glenn.Hanes@ars.usda.gov), Bernard Misof (ZFMK, Center for Molecular Biodiversity Research, Bonn, Germany; b.misof.zfmk@uni-bonn.de), Monica Muñoz-Torres (Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA; mcmunozt@lbl.gov), Oliver Niehuis (ZFMK, Center for Molecular Biodiversity Research, Bonn, Germany; o.niehuis.zfmk@uni-bonn.de), Alexie Papanicolaou (CSIRO Ecosystem Sciences, Black Mountain, Australia; alexie.papanicolaou@csiro.au), Michael Pfrender (University of Notre Dame, South Bend, IN; pfrender.1@nd.edu), Monica Poelchau (Department of Biology, Georgetown University, Washington, DC; mpoelchau@gmail.com), Mary Purcell-Miramontes (USDA, National Institute of Food and Agriculture, Washington, DC; mpurcell@nifa.usda.gov), Hugh M. Robertson (University of Illinois, Urbana-Champaign, IL; hughrobe@life.uiuc.edu), Oliver Ryder (Institute for Conservation Research, San Diego Zoo, San Diego, CA; oryder@sandiegozoo.org), Denis Tagu (NRA-UMR 1349 IGEPP, Rennes, France; denis.tagu@rennes.inra.fr), Tatiana Torres (University of São Paulo, São Paulo, Brazil; tttorres@ib.usp.br), Evgeny Zdobnov (University of Geneva Medical School, Geneva, Switzerland; evgeny.zdobnov@unige.ch), Guojie Zhang (BGI-Shenzhen, China; zhanggj@genomics.org.cn), Xin Zhou (BGI-Shenzhen, China; xinzhou@genomics.org.cn).

References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, et al. 2000. The genome sequence of *Drosophila melanogaster*. Science. 287(5461):2185–2195.

Genome10K Community of Scientists. 2009. Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. Journal of Heredity. 100(6):659–674.

Grbic M, Van Leeuwen T, Clark RM, Rombauts S, Rouzé P, Grbić V, Osborne EJ, Dermauw W, Ngoc PC, Ortego F, et al. 2011. The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. Nature. 479(7374):487–492.

Gregory TR. 2013. Animal Genome Size Database. Available from: http://www.genomesize.com

Received May 5, 2013; First decision June 24, 2013; Accepted July 15, 2013.

Corresponding Editor: José Lopez